# Conceptualisation et exploitation d'une base de données culturelles via des techniques du Web sémantique

---

| | |
|---|---|
| **Formation** | **Master 2 Recherche en Informatique** |
| **Projet de rattachement** | Knowledge and Image analysis for Decision – KID, LGI2P/EMA |
| **Laboratoire d'accueil** | LGI2P – Ecole des mines d'Alès<br>Parc scientifique et technique Georges BESSE<br>30 035 Nîmes cedex – lieu d'accueil Nîmes |
| **Equipe d'encadrement** | Sylvie Ranwez (HDR, LGI2P),<br>Sébastien Harispe (Dr, LGI2P) |
| **Collaboration** | Labo² – Médiathèque du Carré d'Art, Nîmes |
| **Mots-clés** | Indexation conceptuelle, Ontologies, Bases de connaissances, Web sémantique, Recherche d'Information. |
| **Rémunération** | 500,51 €/mois (pendant la période de stage "recherche") |

## Contexte et expression du besoin

Dans de multiples domaines (industrie, santé, culture, etc.), les nouveaux paradigmes des *données ouvertes* et *liées* catalysent la publication de données libres et structurées sur le Web – et contribuent à l'émergence du Web sémantique. Ces changements de pratique ouvrent de nouvelles perspectives et laissent imaginer de nouveaux usages des données. Ils semblent particulièrement prometteurs pour la médiation culturelle en permettant, par exemple, de croiser l'exploitation de données spécifiques aux institutions culturelles avec des connaissances caractérisées sémantiquement et accessibles via les réseaux. Ainsi, jusqu'à récemment, l'accès à un produit culturel était contraint par sa présence physique à un endroit donné, avec un médiateur culturel pour le conseiller, le retrouver ou l'interpréter. Ce n'est plus le cas aujourd'hui où l'accès à ce produit, partiellement ou intégralement numérisé, peut être facilité par des approches innovantes basées sur une analyse de la connaissance des œuvres et de leur contexte. Cette connaissance peut être ambiguë (descriptifs, avis de pairs, critiques) ou bien, à l'image des données liées, être caractérisée par une sémantique maîtrisée et être ainsi directement exploitée par l'outil informatique. Comment dans ces conditions, valoriser un patrimoine culturel et tirer profit des nouvelles technologies pour accompagner au mieux l'usager de produits culturels ? Comment faciliter ses recherches, imaginer ses parcours thématiques, et par exemple lui proposer une sélection pertinente et argumentée d'œuvres en fonction de différents critères (centres d'intérêt, actualité, exposition temporaire dans les environs…) ? C'est une des problématiques que s'est posée la médiathèque du Carré d'Art de Nîmes, via sa structure Labo² (prononcer Labo carré[1]) et qui motive la proposition de ce stage en collaboration avec le laboratoire de recherche LGI2P de l'école des mines d'Alès.

---

[1] Créé en 2012, dans la continuité d'initiatives innovantes que la médiathèque du Carré d'art mène depuis une dizaine d'années, Labo² favorise l'émergence de nouveaux usages du numérique en s'appuyant sur la créativité d'acteurs venant d'univers professionnels différents : artistes, développeurs, associations, entreprises, professionnels des secteurs culturels et éducatifs, publics. Il a reçu en 2012 le label « Bibliothèque numérique de référence » du Ministère de la Culture. Un complément est ajouté en fin de ce document.

Ce stage se situe au carrefour de plusieurs problématiques :

1. La définition d'une stratégie **d'Ingénierie des Connaissances** pour l'utilisation des **technologies du Web sémantique** et des **données liées** en vue d'amener la **recherche** et la **recommandation** de produits culturels. Cette vaste problématique fait référence à l'étude de la caractérisation d'une œuvre en Ingénierie des Connaissances – e.g. identification des types de raisonnement souhaités, de l'expressivité des langages utilisés, des ontologies existantes, des bases de connaissances qu'il est pertinent d'utiliser. Ces travaux doivent permettre la définition d'une base de connaissances adaptée aux besoins du Labo² et de la bibliothèque.

2. L'identification de **liens entre du contenu** textuel ambigu associé à une œuvre, (e.g. descriptif) et des **bases de connaissances** existantes. On fait référence ici à la problématique de *désambiguïsation d'entités nommées*, i.e. comment faire le lien entre une occurrence d'une chaîne de caractères (« Kepler ») et un concept (le télescope spatial ou le scientifique) ?, S'en suit la problématique d'*indexation conceptuelle* – comment résumer un texte de façon à en distinguer les thématiques importantes ? – et comment intégrer cette information dans une base de connaissances ?

3. L'utilisation de la base de connaissances dans une stratégie de **Recherche d'Information et de Recommandation** – sauf volonté forte, il ne sera pas demandé au stagiaire d'étudier les aspects algorithmiques associés à ces problématiques.

## Organisation du stage de recherche

<u>Etat de l'art</u>

Il se décomposera en deux parties :

1. Après une première sensibilisation à la problématique d'indexation (e.g [1]), l'état de l'art analysera les contributions permettant d'identifier la définition d'une stratégie d'Ingénierie des Connaissances à mettre en œuvre (voir objectif 1 ci-dessus). La littérature analysée sera par exemple celle produite à l'occasion de rencontres comme « Library Linked Data: Let's make it happen! »[2]. Il sera naturellement demandé au stagiaire de rentrer en contact avec les acteurs français qui ont une forte expérience dans le domaine, e.g. Bnf. L'état de l'art s'intéressera donc aux méthodologies et ressources termino-ontologiques utilisées par les bibliothèques et autres organismes culturels pour faciliter l'accès à des œuvres (e.g. bnf, [3], [4])

2. La deuxième partie de l'état de l'art s'intéressera tout particulièrement aux techniques de désambiguïsation d'entités nommées [5][6][7], qui permettent de faire le lien entre les métadonnées et les descriptifs des ressources (e.g. œuvres culturelles) et des bases de connaissances existantes (e.g. DBpedia, Yago2). Le stagiaire s'intéressera ensuite à la conceptualisation de textes – comment résumer les liens entres les entités nommées et le texte analysé, et comment stocker cette indexation dans la base de connaissances ?

<u>Recherche théorique et appliquée</u>

La deuxième partie de ce stage, effectuée au sein du laboratoire LGI2P localisé à Nîmes, concernera la conceptualisation du catalogue de la médiathèque[2] du Carré d'Art. Durant ce stage de recherche, le candidat devra définir la stratégie d'Ingénierie des Connaissances qu'il convient d'adopter et proposer l'approche permettant de construire la base de connaissances – en prenant en compte les aspects relatifs à l'indexation conceptuelle et les objectifs de Recherche d'Information et de Recommandation

[2] http://cat-bib.nimes.fr/index.html

introduits. Le développement d'un prototype est envisagé en collaboration avec des membres de l'équipe[3].

## Compétences

Ce stage s'adresse à un(e) étudiant(e) en Informatique en 2e année de Master. Le/la candidat(e) devra avoir des affinités avec la problématique associée à la médiation culturelle et être fortement intéressé(e) par l'Ingénierie des Connaissances et les technologies du Web sémantique. Une bonne maîtrise des spécifications RDF(S)/OWL/SPARQL/SKOS serait un plus. A noter qu'un bon niveau en anglais est requis pour mener efficacement l'étude bibliographique. Dans l'idéal, et si l'étudiant(e) le souhaite, il est envisagé de développer un prototype permettant une mise en pratique de l'approche proposée en réponse aux besoins définis – des données issues du Carré d'Art sont disponibles. Dans ce cas, les développements de la partie serveur, en collaboration avec les membres de l'équipe, seront effectués en langage Java – une connaissance de la librairie Jena ou Sesame serait ici un plus.

## Contacts


| | | |
|---|---|---|
| Sylvie Ranwez | sylvie.ranwez@mines-ales.fr | 04 66 38 70 44 |
| Sébastien Harispe | sebastien.harispe@mines-ales.fr | 04 66 38 70 36 |


## Labo²

Le Labo² travaille au développement des usages artistiques et culturels du numérique à la bibliothèque Carré d'Art et dans les établissements publics/privés partenaires. Il s'appuie sur les compétences développées par les artistes numériques pour imaginer, avec les publics, les médiateurs, les chercheurs des dispositifs de médiation culturels innovants.
Les objectifs de Labo² sont de plusieurs ordres : stimuler l'innovation technologique et sociale, promouvoir les arts et cultures numériques auprès de tous, favoriser l'émergence d'un écosystème créatif sur Nîmes et son agglomération par l'accompagnement et le développement des usages numériques innovants. Labo² fait partie du projet de service de la bibliothèque labellisé par le ministère de la culture Bibliothèque Numérique de Référence (B.N.R).

## Références


[1] Muriel Amar. L'indexation aujourd'hui http://www2.cndp.fr/archivage/valid/68661/68661-10219-12814.pdf

[2] Collectif. Session " Library Linked Data: Let's make it happen!", IFLA, Lyon, 2014. http://ifla2014-satdata.bnf.fr/program.html

[3] Patrick Le Boeuf. « ... Qui se pavane et s'agite son heure sur scène et qu'ensuite on n'entend plus... » : les éléments à prendre en compte dans un futur modèle conceptuel du spectacle vivant et de l'information relative à ses archives Intervention donnée lors de la journée d'étude « De la conception à la survie : comment documenter et conserver les productions du spectacle multimédia ? », Centre de documentation de la musique contemporaine, 13 janvier 2006, Paris. http://www.cidoc-crm.org/docs/2006_LeBoeuf_fr.pdf

[4] The CIDOC Conceptual Reference Model http://www.cidoc-crm.org/


[3] http://kidknowledge.wp.mines-telecom.fr/

[5] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, Gerhard Weikum. Robust Disambiguation of Named Entities in Text, In: Conference on Empirical Methods in Natural Language Processing, p. 782–792, Edinburgh, Scotland, 2011

[6] Pablo N. Mendes, Max Jakob, Andrés García-Silva and Christian Bizer. DBpedia Spotlight: Shedding Light on the Web of Documents. Proceedings of the 7th International Conference on Semantic Systems (I-Semantics). Graz, Austria, 7–9 September 2011.

[7] Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. A framework for benchmarking entity-annotation systems. In Proceedings of the 22nd international conference on World Wide Web (WWW '13). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 249-260, 2013.